(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 321 848 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
25.06.2003 Bulletin 2003/26

(51) Int Cl.[7]: **G06F 3/06**

(21) Application number: **02258904.8**

(22) Date of filing: **20.12.2002**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SI SK TR
Designated Extension States:
AL LT LV MK RO

(30) Priority: **21.12.2001 US 27457**

(71) Applicant: **Network Appliance, Inc.**
Sunnyvale, California 94089 (US)

(72) Inventors:
- **Coatney, Susan M.**
  Cupertino, California 95014 (US)
- **Rowe, Alan L.**
  San Jose, California 95123 (US)
- **Aster, Radek**
  Campbell, California 95008 (US)
- **Sarma, Joydeep Sen**
  Redwood City, California 94062 (US)

(74) Representative: **Collins, John David**
Marks & Clerk,
57-60 Lincoln's Inn Fields
London WC2A 3LS (GB)

(54) System and method of implementing disk ownership in networked storage

(57) A system and method for disk ownership in a network storage system. Each disk has two ownership attributes set to show that a particular file server owns the disk. In a preferred embodiment the first ownership attribute is the serial number of the file server being written to a specific location on each disk and the second ownership attribute is setting a SCSI-3 persistent reservation. In a system utilizing this disk ownership method, multiple file servers can read data from a given disk, but only the file server that owns a particular disk can write data to the disk.

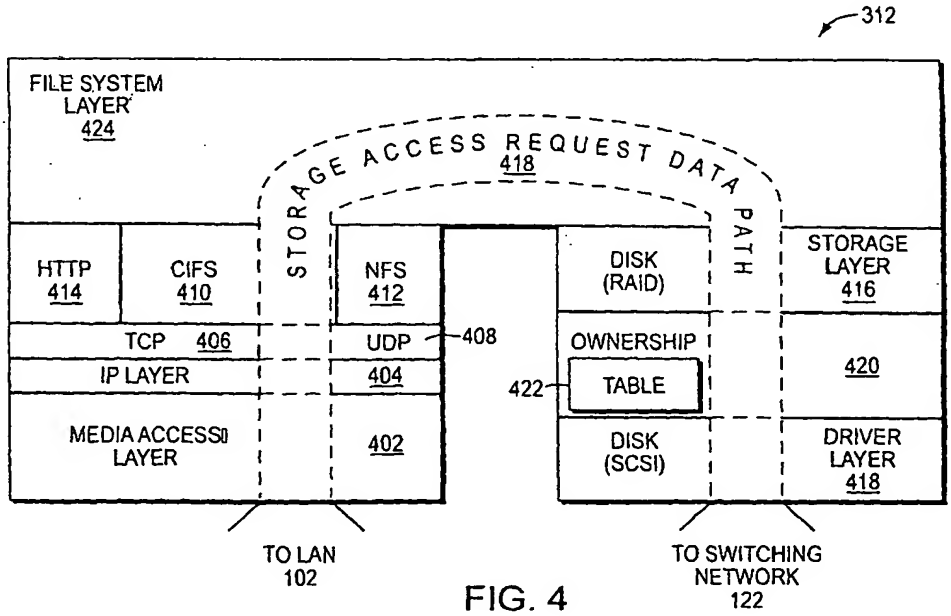

FIG. 4

EP 1 321 848 A2

This capability can be utilized to generate redundant data pathways to a disk.

[0009] Each of the devices attached to the LAN include an appropriate conventional network interface arrangement (not shown) for communicating over the LAN using desired communication protocol such as the well-known Transport Control Protocol/Internet Protocol (TCP/IP), User Datagram Protocol (UDP), Hypertext Transfer Protocol (HTTP), or Simple Network Management Protocol (SNMP).

[0010] One prior implementation of a storage system involves the use of switch zoning. Instead of the filer being directly connected to the fibre channel loop, the filer would be connected to a fibre channel switch, which would then be connected to a plurality of fibre channel loops. Switch zoning is accomplished within the fibre channel switches by manually associating ports of the switch. This association with, and among, the ports would allow a filer connected to a port associated with a port connected to a fibre channel loop containing disks to "see" the disks within that loop. That is, the disks are visible to that port. However, a disadvantage of the switch zoning methodology was that a filer could only see what was within its zone. A zone is defined as all devices that are connected to ports associated with the port to which the filer was connected. Another noted disadvantage of this switch zoning method is that if zoning needs to be modified, an interruption of service occurs as the switches must be taken off-line to modify zoning. Any device attached to one particular zone can only be owned by another device within that zone. It is possible to have multiple filers within a single zone; however, ownership issues then arise as to the disks within that zone.

[0011] The need, thus, arises for a technique for a filer to determine which disks it owns other than through a hardware mechanism and zoning contained within a switch. This disk ownership in a networked storage methodology would permit easier scalability of networked storage solutions.

Summary of the Invention

[0012] One aspect of the invention overcomes the disadvantages of the prior art by providing a system and method of implementing disk ownership by respective file servers without the need for direct physical connection or switch zoning within fibre channel (or other) switches. A two-part ownership identification system and method is defined. The first part of this ownership method is the writing of ownership information to a predetermined area of each disk. Within the system, this ownership information acts as the definitive ownership attribute. The second part of the ownership method is the setting of a SCSI-3 persistent reservation to allow only the disk owner to write to the disk. This use of a SCSI-3 persistent reservation allows other filers to read the ownership information from the disks. It should be noted that other forms of persistent reservations can be used in accordance with the invention. For example, if a SCSI level 4 command set is generated that includes persistent reservations operating like those contained within the SCSI-3 command, these new reservations are expressly contemplated to be used in accordance with the invention.

[0013] By utilizing this ownership system and method, any number of file servers connected to a switching network can read from, but not write to, all of the disks connected to the switching network. In general, this novel ownership system and method enables any number of file servers to be connected to one or more switches organized as a switching fabric with each file server being able to read data from all of the disks connected to the switching fabric. Only the file server that presently owns a particular disk can write to a given disk.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The above and further advantages of the invention may be better understood by referring to the following description in conjunction with the accompanying drawings in which like reference numerals indicate identical or functionally similar elements:

Fig. 1, already described, is a schematic block diagram of a network environment showing the prior art of a filer directly connected to fibre channel loop;

Fig. 2 is a schematic block diagram of a network environment including various network devices including exemplary file servers and associated volumes;

Fig. 3 is a schematic block diagram of an exemplary storage appliance in accordance with Fig. 2;

Fig. 4 is a schematic block diagram of a storage operating system for use with the exemplary file server of Fig. 3 according to an embodiment of this invention;

Fig. 5 is a block diagram of an ownership table maintained by the ownership layer of the storage operating system of Fig. 4 in accordance with an embodiment of this invention; and

Fig. 6 is a flow chart detailing the steps performed by the storage operating system upon boot up to obtain ownership information of all disks connected to fibre channel switches connected to the individual filer.

purpose computer's microprocessor identification number, the file server's media access code (MAC) address, etc.

[0023] In the illustrative embodiment, the memory 304 may have storage locations that are addressable by the processor for storing software program code or data structures associated with the present invention. The processor and adapters may, in turn, comprise processing elements and/or logic circuitry configured to execute the software code and manipulate the data structures. The storage operating system 312, portions of which are typically resident in memory and executed by the processing elements, functionally organize a file server by *inter-alia* invoking storage operations in support of a file service implemented by the file server. It will be apparent by those skilled in the art that other processing and memory implementations, including various computer readable media may be used for storing and executing program instructions pertaining to the inventive technique described herein.

[0024] The network adapter 306 comprises the mechanical, electrical and signaling circuitry needed to connect the file server to a client over the computer network, which as described generally above, can comprise a point-to-point connection or a shared medium such as a LAN. A client can be a general-purpose computer configured to execute applications including file system protocols, such as the Network File System (NFS) or the Common Internet File System (CIFS) protocol. Moreover, the client can interact with the file server in accordance with the client/server model of information delivery. The storage adapter cooperates with the storage operating system 312 executing in the file server to access information requested by the client. The information may be stored in a number of storage volumes (Volume 0 and Volume 1) each constructed from an array of physical disks that are organized as RAID groups (RAID GROUPs 1, 2 and 3). The RAID groups include independent physical disks including those storing a striped data and those storing separate parity data. In accordance with a preferred embodiment RAID 4 is used. However, other configurations (e.g., RAID 5) are also contemplated.

[0025] The storage adapter 308 includes input/output interface circuitry that couples to the disks over an I/O interconnect arrangement such as a conventional high-speed/high-performance fibre channel serial link topology. The information is retrieved by the storage adapter, and if necessary, processed by the processor (or the adapter itself) prior to being forwarded over the system bus to the network adapter, where the information is formatted into a packet and returned to the client.

[0026] To facilitate access to the disks, the storage operating system implements a file system that logically organizes the information as a hierarchical structure of directories in files on the disks. Each on-disk file may be implemented as a set of disk blocks configured to store information such as text, whereas the directory may be implemented as a specially formatted file in which other files and directories are stored. In the illustrative embodiment described herein, the storage operating system associated with each volume is preferably the NetApp® Data CNTAP storage operating system available from Network Appliance Inc. of Sunnyvale, California that implements a Write Anywhere File Layout (WAFL) file system. The preferred storage operating system for the exemplary file server is now described briefly. However, it is expressly contemplated that the principles of this invention can be implemented using a variety of alternate storage operating system architectures.

[0027] The host adapter 316, which is connected to the storage adapter of the file server, provides the file server with a unique world wide name, described further below.

C. Storage Operating System and Disk Ownership

[0028] As shown in Fig. 4, the storage operating system 312 comprises a series of software layers including a media access layer 402 of network drivers (e.g., an Ethernet driver). The storage operating system further includes network protocol layers such as the Internet Protocol (IP) layer 404 and its Transport Control Protocol (TCP) layer 406 and a User Datagram Protocol (UDP) layer 408. A file system protocol layer provides multi-protocol data access and, to that end, includes support from the CIFS protocol 410, the Network File System (NFS) protocol 412 and the Hypertext Transfer Protocol (HTTP) protocol 414.

[0029] In addition, the storage operating system 312 includes a disk storage layer 416 that implements a disk storage protocol such as a RAID protocol, and a disk driver layer 418 that implements a disk access protocol such as e.g., a Small Computer System Interface (SCSI) protocol. Included within the disk storage layer 416 is a disk ownership layer 420, which manages the ownership of the disks to their related volumes. Notably, the disk ownership layer includes program instructions for writing the proper ownership information to sector S and to the SCSI reservation tags.

[0030] As used herein, the term "storage operating system" generally refers to the computer-executable code operable on a storage system that implements file system semantics (such as the above-referenced WAFL) and manages data access. In this sense, ONTAP software is an example of such a storage operating system implemented as a microkernel. The storage operating system can also be implemented as an application program operating over a general-purpose operating system, such as UNIX® or Windows NT®, or as a general-purpose operating system with configurable functionality, which is configured for storage applications as described herein.

[0031] Bridging the disk software layers, with the network and file system protocol layers, is a file system layer 424 of the storage operating system. Generally, the

invention can be implemented as a computer program, the present invention encompasses any suitable carrier medium carrying the computer program for input to and execution by a computer. The carrier medium can comprise a transient carrier medium such as a signal e.g. an electrical, optical, microwave, magnetic, electromagnetic or acoustic signal, or a storage medium e.g. a floppy disk, hard disk, optical disk, magnetic tape, or solid state memory device. Additionally, it is expressly contemplated that other devices connected to a network can have ownership of a disk in a network environment. Accordingly, this description is meant to be taken only by way of example and not to otherwise limit the scope of this invention.

Claims

1. A method for a network device to claim ownership of a disk in a network storage system comprising the steps of:

   setting a first ownership attribute on the disk to a state of ownership by network device; and
   setting a second ownership attribute on the disk to a state of ownership by network device.

2. The method of claim 1, wherein one of the first ownership attribute and the second ownership attribute further comprises a small computer system interface level 3 persistent reservation tag.

3. The method of claim 1, wherein one of the first ownership attribute and the second ownership attribute further comprises ownership information written on a predetermined area of the disk.

4. The method of claim 3, wherein the ownership information further comprises a serial number of the network device.

5. The method of any preceding claim, wherein the network device comprises a file server.

6. A method of claiming ownership of a disk by a network device in a network storage system comprising the steps of:

   writing ownership information to a predetermined area of the disk; and
   setting a small computer system interface level 3 persistent reservation tag to a state of network device ownership.

7. The method of claim 6 wherein the ownership information further comprises a serial number of a network device.

8. The method of claim 6 or claim 7, wherein the network device comprises a file server.

9. A network storage system comprising:

   a plurality of network devices;
   one or more switches, each network device connected to at least one of the one or more switch; and
   a plurality of disks having a first ownership attribute and a second ownership attribute, each disk connected to at least one of the plurality of switches.

10. The network storage system of claim 9, wherein the first ownership attribute further comprises ownership information written on a predetermined area of the disk.

11. The network storage system of claim 9 or claim 10, wherein the second ownership attribute further comprises a small computer system interface level 3 persistent reservation tag.

12. The networked storage system of claim 11, wherein each disk that is owned by the network device has the small computer system interface level 3 persistent reservation set such that only the network device may write to the disk.

13. The network storage system of claim 10, wherein the ownership information further comprises of a serial number of the network device that owns that particular disk.

14. The network storage system of any one of claims 9 to 13, wherein each of the plurality of file servers can read data from each of the plurality of disks.

15. The network storage system of any one of claims 9 to 14, wherein only a network device that owns one of the plurality of disks can write data to the one disk.

16. The network storage system of any one of claims 9 to 15, wherein the network devices comprise file servers.

17. A network storage system comprising:

   one or more switches;
   a plurality of disks; and
   a plurality of network devices, each of the network devices including means for claiming ownership of one of the plurality of disks in the network storage system.

18. The network storage system of claim 17, wherein the means for claiming ownership further compris-

EP 1 321 848 A2

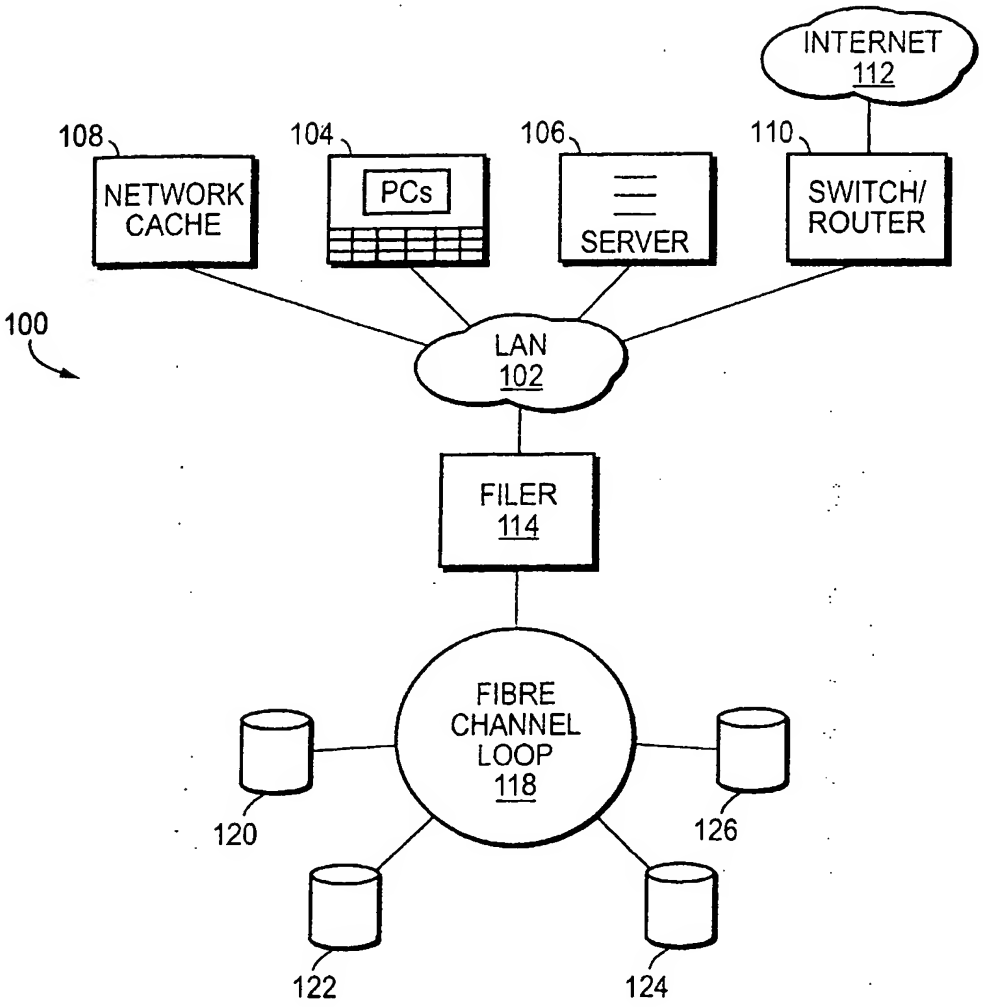

FIG. 1
(PRIOR ART)

EP 1 321 848 A2

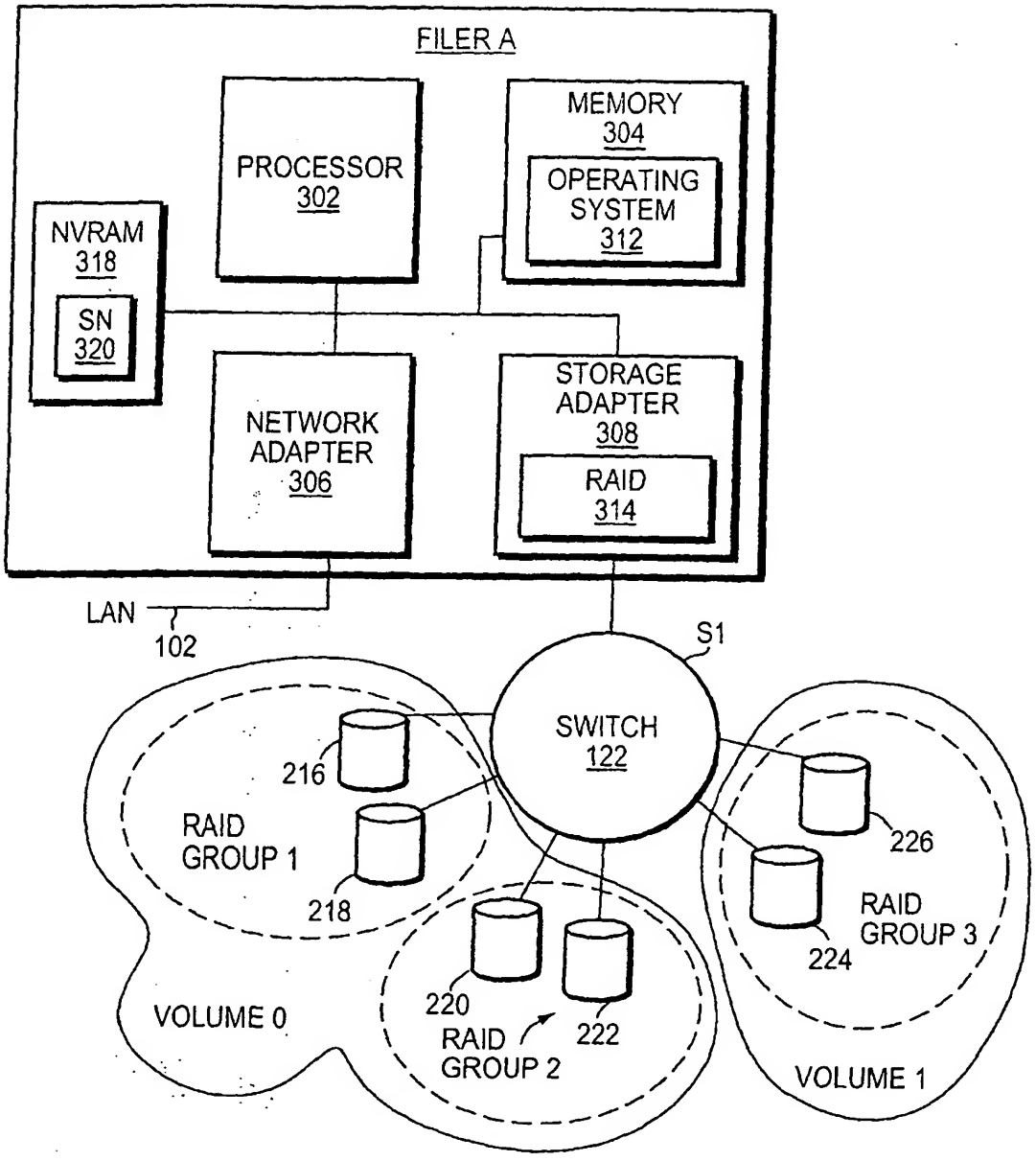

FIG. 3

EP 1 321 848 A2

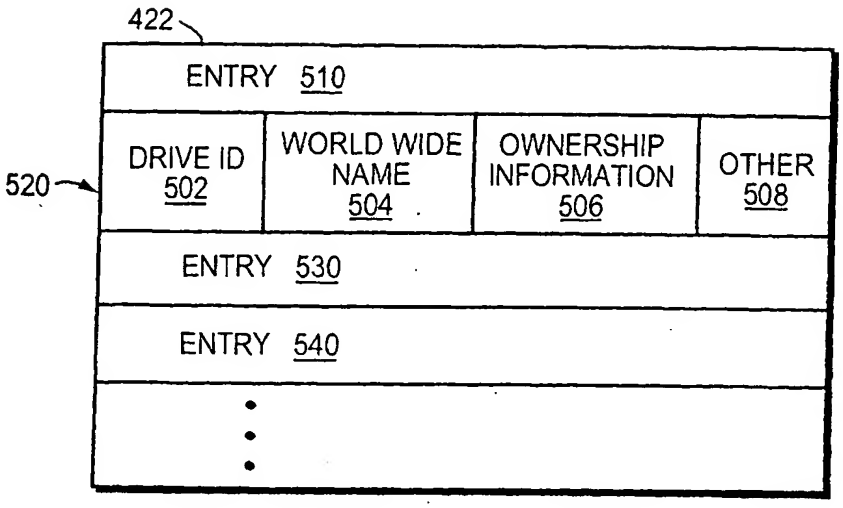

FIG.5